**HOUSE PRICE PREDICTION**

***A Project***

***Submitted in partial fulfillment of the requirements for***

***the award of the Degree of***

**BACHELOR OF COMPUTER APPLICATION**

**By**

**SOUMYADEEP NANDI**

**ROLL NO-143 AND REGISTRATION NO-203661001210131**

**TWINKLE ROY**

**ROLL NO-182 AND REGISTRATION NO-203661001210159**

**SREEJA BISWAS**

**ROLL NO-152 AND REGISTRATION NO-203661001210164**

**SHREYA MAJUMDER**

**ROLL NO-125 AND REGISTRATION NO-203661001210127**

**SHUBHAMITA SAHA**

**ROLL NO-126 AND REGISTRATION NO-203661001210120**



**DEPARTMENT OF COMPUTER APPLICATION**

**INSTITUTE OF ENGINEERING & MANAGEMENT**

**2023**

**DECLARATION CERTIFICATE**

This is to certify that the work presented in the thesis entitled **“HOUSE PRICE PREDICTION”** in partial fulfillment of the requirement for the award of degree of **Bachelor of Computer Application** of Institute of Engineering & Management is an authentic work carried out under my supervision and guidance.

To the best of my knowledge the content of this thesis does not form a basis for the award of any previous Degree to anyone else.

Date: 02/05/23 Faculty Name: Manjima Saha

Dept. of Computer Application

Institute of Engineering & Management

Dr. Abhishek Bhattacharya

Head of the Department

Dept. of Computer Application and Science

Institute of Engineering & Management

**CERTIFICATE OF APPROVAL**

The foregoing thesis entitled **“HOUSE PRICE PREDICTION”** is hereby approved as a creditable study of research topic and has been presented in satisfactory manner to warrant its acceptance as prerequisite to the degree for which it has been submitted.

It is understood that by this approval, the undersigned do not necessarily endorse any conclusion drawn or opinion expressed therein, but approve the thesis for the purpose for which it is submitted.

**(Internal Examiner) (External Examiner)**

**Acknowledgements**

We would like to express our special thanks of gratitude to our Guide **Prof. Manjima Saha** who helped us a lot in this project, her valuable suggestions helped us to solve tough challenges and without her help this project could not have been completed in time. A special thanks to our Head of Department **Prof. Abhishek Bhattacharya** who gave us the golden opportunity to do this wonderful project on the topic **“HOUSE PRICE PREDICTION”**, which helped us to gain a significant knowledge in the aforesaid subjects. Secondly, we would like to thank our friends who helped us a lot in finalizing this project within the given time frame.

**Name of Student: Soumyadeep Nandi**

**Enrollment Number: 12020004006142**

**Name of Student: Twinkle Roy**

**Enrollment Number: 12020004006182**

**Name of Student: Sreeja Biswas**

**Enrollment Number: 12020004006151**

**Name of Student: Shreya Majumder**

**Enrollment Number: 12020004006124**

**Name of Student: Shubhamita Saha**

**Enrollment Number: 12020004006125**

**Contents**

**Abstract** ………………………………………………………………………………………………………..vi

**Chapter 1**

1.1 Introduction *………………………………………………………………………….………………………*7

**Chapter 2**

2.1 Background Studies*……………………………………………………………………………….……*.8

2.2 Literature Survey ….……*……………………………………………………………………………….9*

**Chapter 3**

3.1 Proposed Methodology…………………………………….*…………………………………….*……10

**Chapter 4**

4.1 Experimental Dataset…………………………………………………………………………………...12

**Chapter 5**

5.1 Results and Discussion*…………………………………………………………………………………*.15

**Chapter 6**

6.1 Conclusions …………………………………………..……………………………………………………18

6.2 Future Work …………………………………………….…………………………………………………18

**Abstract**

Machine learning has become more important over the past few years in a variety of fields, including image detection, spam recognition, speech commands, product suggestions, and medical diagnosis. Today’s sophisticated machine learning algorithms are assisting us in enhancing security warnings, enhancing public safety, and advancing medical research. Moreover, machine learning technologies offer safer automotive systems and improved customer service. In this study, we concentrate on applying machine learning techniques to forecast future home prices. We examined and investigated different prediction systems in order to choose the optimal one. Therefore, we selected lasso regression as our model because of its versatility and probabilistic model selection process. Our findings demonstrate the effectiveness of our method for solving the issue and its capacity to generate forecasts that are on par with those from other models for predicting housing costs. The establishment of home value indexes, which in turn promotes real estate policies and plans, is another contribution made by this study. To develop a housing cost prediction model, we used machine learning techniques as a research strategy. To assess the performance of different machine learning algorithm models, including gradient boosting boost, lasso regression, and neural random forest, we tested and compared them. According to our research, the lasso regression method predicts housing costs more accurately than other models on a constant basis. Overall, our study advises making more educated decisions about a home’s valuation with the use of the lasso regression method for sellers or real estate brokers. We can create dependable and accurate housing cost prediction models with the aid of machine learning, which will ultimately help the real estate market as a whole.

**Chapter 1**

**1.1 Introduction**

Machine learning is a branch of artificial intelligence (AI) that focuses on using technology and algorithms to extract information that is usable from data, especially in the case of enormous data where manual processing is impractical. In contrast to mathematical approaches, machine learning makes use of algorithms that give computers the ability to learn to solve issues algorithmically. Supervised and unsupervised machine learning fall into two broad types. Unsupervised learning focuses on finding connections and hidden patterns within data, while supervised learning includes training a program on a specified set of data in order to predict fresh data Today’s real-world applications employ a range of machine learning algorithms, each performing differently depending on the situation. This thesis compares how well artificial neural networks and regression algorithms perform at forecasting values from a given dataset. Many other papers used several other algorithms in order to predict the accuracy and we have taken the dataset from Kaggle kc\_house\_data. Since the prediction of house prices relies on a wide range of criteria that can vary in price depending on location and other circumstances, house prices will specifically be used as a benchmark for comparison. We added new variables to the local dataset in order to examine the relationship between these variables and the sale price. We came to the conclusion that our model is more efficient to predict the price of the house compared to any other model.

**Chapter 2**

**2.1 Background Studies**

In our ecosystem, real estate has the lowest level of transparency. As opposed to being forecasted based on valuation, estate prices can get exaggerated and fluctuate daily. Predicting property prices using actual factors is the central tenet of our exploratory approach. Python programming languages are employed to carry out this work. In order to forecast home prices with greater accuracy and superior results, this study is proposed. It's one of the high domains to apply machine learning techniques on how to improve and prevision expenses with high delicacy. The megacity that we are talking about is among the most significant locations in King County, Washington, or any other nearby region. Houses are 138 more valuable than average because it is a fashionable areas that people want to live in. More nearby this location will be more precious. Machine learning has emerged as a crucial vaticination method as a result of the expanding trend towards Big Data since it can forecast property values more directly based on their characteristics, regardless of the data from previous periods. Numerous research looked at this issue and shown the effectiveness of the machine learning strategy. We were able to verify the effectiveness of each strategy by applying several styles to this dataset. Algorithms for machine learning are utilized in both exploration and numerous practical activities. Large structured and georeferenced datasets are now more widely available as a result of digital technology, making it easier to utilize these algorithms to analyze data, spot trends, and predict outcomes that support decision-making.

**2.2 Literature Survey**

A number of computer science methods are used to anticipate home prices, which is a broad field. For example, machine learning, linear regression, decision trees, deep learning, fuzzy logic, and linear performance pricing.

This paper's main objective is to investigate various machine learning techniques that can improve the training of machine learning models that forecast house prices. The pricing of assets, particularly housing, and their relationship to monetary fluctuations are topics of increasing concern in academic and strategic circles as a result of the current financial crisis. The cost of a home is determined by a variety of elements, including the number of bedrooms, bathrooms, kitchen, balcony, and location of the house. Historically, real estate firms have depended on manual forecasts created by a management team to ascertain the value of a piece of real estate, which frequently causes losses for both purchasers and sellers. Due to these manual forecasts, losses in the past were up to 25% higher than they are now. In order to avoid potential losses from manual estimates, the main goal of this work is to create an AI system that can reliably anticipate housing values.

Similar experiments have been carried out by researchers from other states, and the results are encouraging, with the majority of them reaching an output accuracy of above 75%. We shall describe various real-time experimental research projects carried out by scientists in different nations in the part that follows, which may be used as examples for our study.

The dataset used in Anand G. Rawool and Dattatray V. Rogye's work that dataset is split into two sections: testing and training. Data is utilized for training purposes 80% of the time and testing purposes 20% of the time. The target variable is included in the training set. Several machine learning methods are used to train the model, however, Random forest regressions produce the best predictions. The Python libraries NumPy and Pandas were utilized to implement the algorithms. Anirudh Kaushal and Achyut Shankar utilized the multivariate linear regression model to make the prediction in another research based on machine learning. It is also contrasted with other machine learning models including the decision tree regressor, Ridge, RidgeCV, Lasso, and LassoCV. With an accuracy rate of 84.5%, multivariate linear regression, and LassoCV perform best. And the other paper is based on the Machine Learning of G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch. Srinivasulu has used XGBoost, lasso regression and neural systems to look at their order precision execution model. Gradient boosting regression performs the best with 91.2% accuracy.

These are some examples that we got doing our survey for the research work which we’re going to do in this paper. Here our research discussion will be forecasted to all.

**Chapter 3**

**3.1 Proposed Methodology**

Now that the data is cleaned of noisy data and pre-processed, we can finally use it for prediction. In this section, we will discuss the algorithm we have used for the prediction of the property price and how the training and testing of the model will take place.

Used algorithm:

Mainly using different types of supervised algorithms is the best choice to predict for house price prediction.

**Supervised algorithm:**

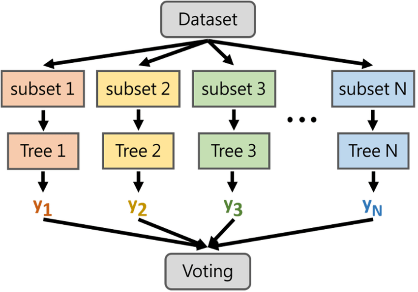
Supervised learning is also known as supervised machine learning is a subcategory of machine learning and artificial intelligence.it is defined by its use of labelled datasets to train algorithms that classify data or predict outcomes accurately**.** Here we’ve used different regression to find the accurate prediction:- 1) Linear regression 2) Lasso regression 3) Gradient boosting regression 4) Random forest regression

1. **Linear regression:** The algorithm that we have selected is linear regression, where the value for the dependent variable is calculated using multiple independent variables. The value of the variable which is to be predicted depends on its strength of relationship with the other independent variables. This factor is called correlation.

Y = mx + c [y is the dependent variable, x is the independent variable, and m is the intersection factor]

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

1. **Lasso regression:** Lasso was originally formulated for linear regression models. This simple case reveals a substantial amount about the estimator. These include its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that the coefficient estimates do not need to be unique if covariates are collinear. Though originally defined for linear regression, lasso regularization is easily extended to other statistical models including generalized linear models, generalized estimating equations, and proportional hazards models.
2. **Gradient boosting regression:**Gradient boosting works by multiple residual models together. The error value (difference between the actual and predicted value) and the X value are taken as input at each start of the residual model. Then after calculating the prediction, its error value is fed into the next residual model. This process goes on until the error becomes minimum. Gradient boosting is a [machine learning](https://en.wikipedia.org/wiki/Machine_learning) technique used in [regression](https://en.wikipedia.org/wiki/Regression_(machine_learning)) and [classification](https://en.wikipedia.org/wiki/Classification_(machine_learning)) tasks, among others. It gives a prediction model in the form of an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning) of weak prediction models, which are typically [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning). When a decision tree is a weak learner, the resulting algorithm is called a gradient-boosted tree; it usually outperforms [random forest](https://en.wikipedia.org/wiki/Random_forest).  A gradient-boosted trees model is built in a stage-wise fashion as in other [boosting](https://en.wikipedia.org/wiki/Boosting_(machine_learning)) methods, but it generalizes the other methods by allowing optimization of an arbitrary [differentiable](https://en.wikipedia.org/wiki/Differentiable_function) [loss function](https://en.wikipedia.org/wiki/Loss_function).
3. **Random forest regression:** The random forest algorithm works by taking different decision trees together and combining them. Next predictions are made for each and averaging is done based on these predictions. The prediction is made based on the most accurate results. The following flowchart describes the working of the random forest algorithm.



**Chapter 4**

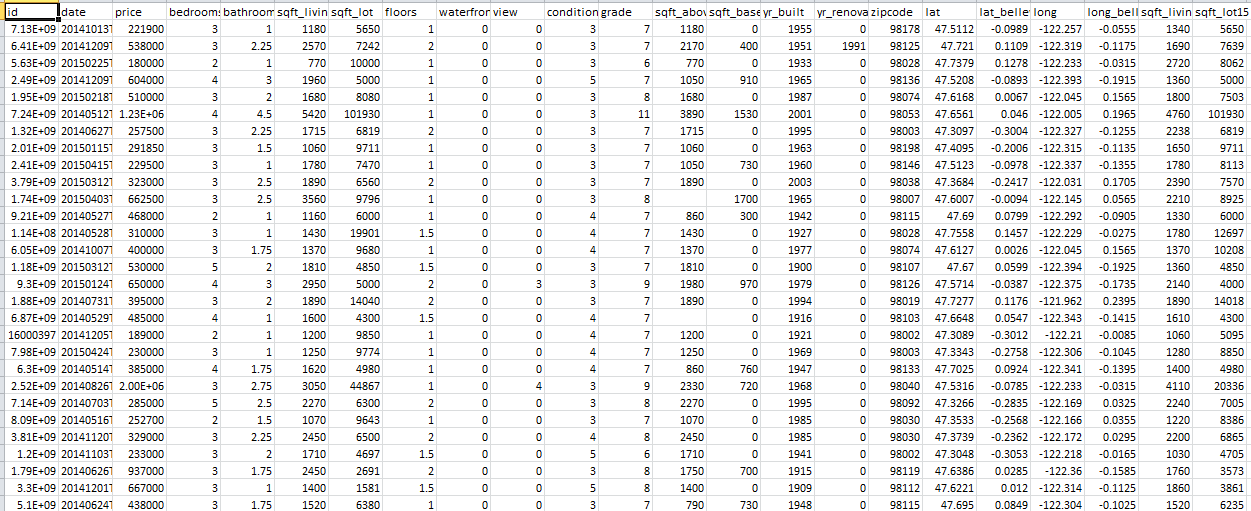
**4.1 Experimental Dataset**

**Data collection:**

We have collected datasets of house price and all other attributes that are applicable from online website. We have downloaded the .csv files in which information was present.

**Data Pre-processing:**

Data pre-processing includes data collection, data formation, and data modification. We’ve collected the exact dataset that they’ve used in their paper from kaggle. We add some new attribute in the dataset kc\_house\_data. Which consist of 23 attribute and 21614 columns. This dataset is a collection of the data information of all the houses in us and surrounding king county and it’s all other cities. After collecting and modifying first thing we checked for any missing or null data values. Then that clean data was uploaded for the next procedure. That dataset is given following.



**Experimental data analysis:**

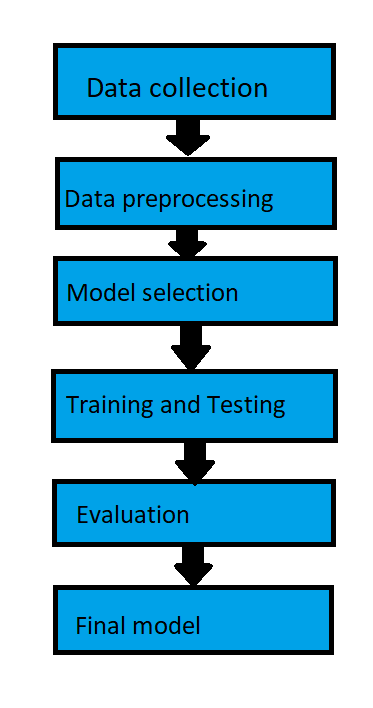
Data analyzing is a very essential step to perform analytically the dataset that we’ve collected and cleaned that was not enough to predict best result as we see in other papers. Some machine learning that they used  was getting more accurate result than other algorithms. So we add some new attribute in the dataset and tried to make it efficient enough to predict best result using other algorithms. We consider a nearby city Bellevue and its longitude and latitude 47.61010 N, 122.20150W. We tried to experiment with adding a few new attributes and doing subs traction operation with the entire element in the longitude and latitude column with the location of that nearest state which is mentioned above. Doing this experiment we got the best results accordingly.

**Model selection:**

The next most important step is selecting suitable models. In this case, regression algorithms are the most suitable that we used for prediction.to perform on regression algorithm model data need to be pre-processed and cleaned otherwise different can occur.  We select a few machine learning algorithms like linear regression algorithm, lasso regression algorithm, random forest regression algorithm, and gradient boosting algorithm.

**Training and testing of dataset:**

The data sets were divided such that x\_train is used to train the model with corresponding x\_test values and some y\_train is kept reserved for testing. The model was tested with y\_train and stored in y\_predict. Both y\_train and y\_predict were compared.



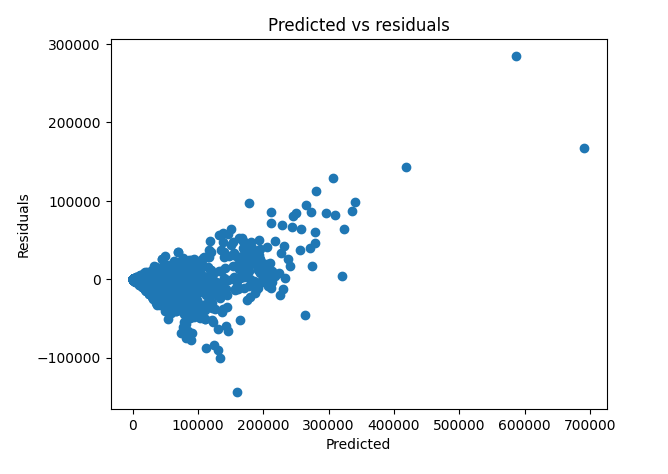
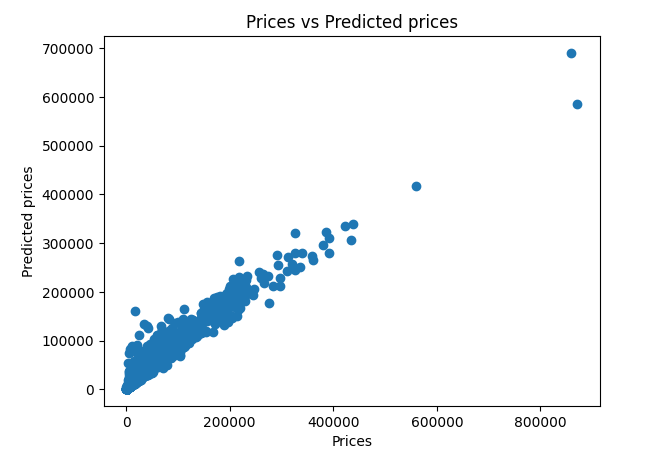
**Chapter 5**

**5.1 Result and Discussions**

Different model gives different result among that random forest model gives the most accurate result.

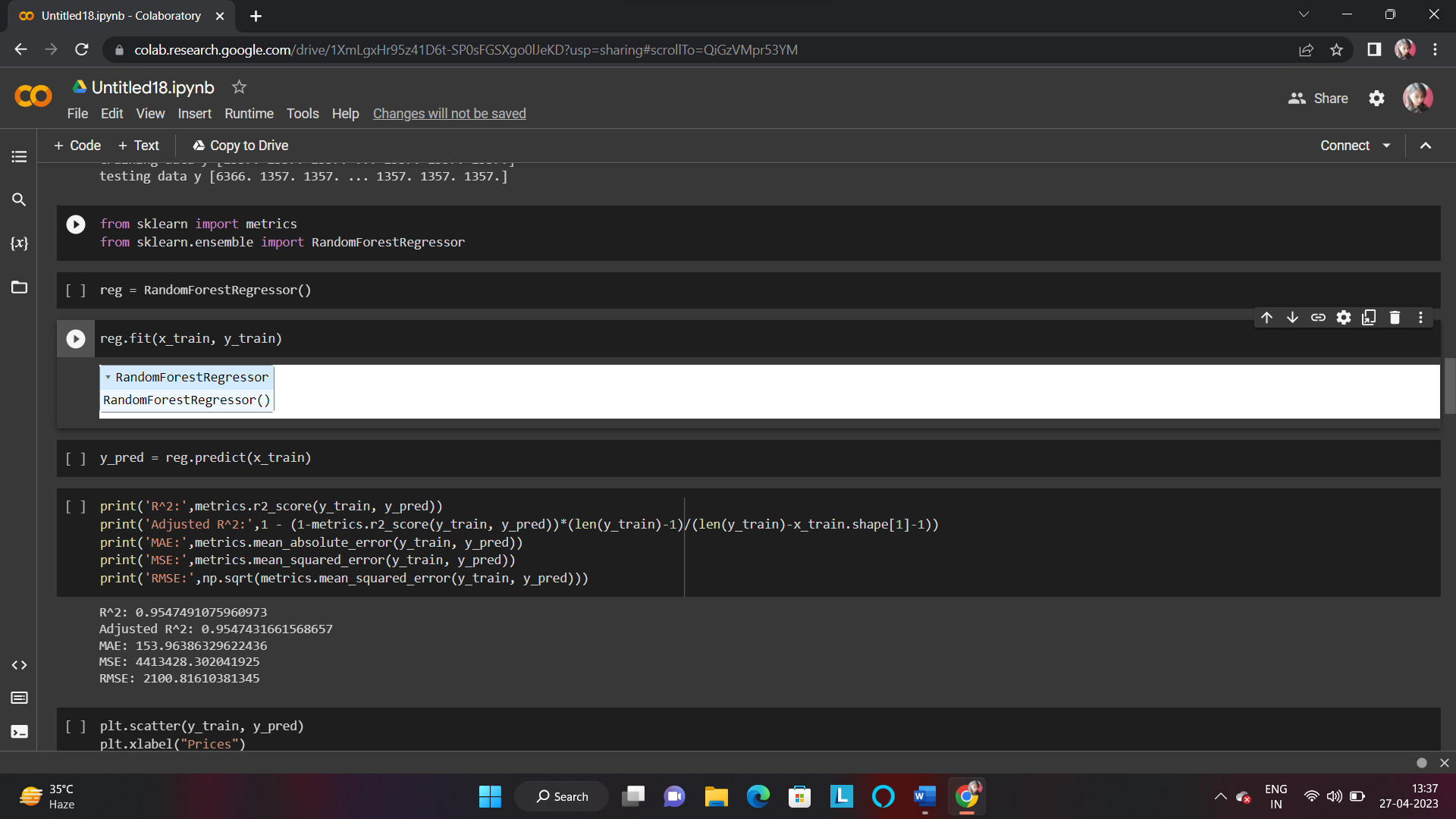
|  |  |
| --- | --- |
| Model | R2 adjust matrix |
| Linear regression | 0.55 |
| Lasso regression | 0.61 |
| Random forest regression | 0.95 |
| Gradient boosting | 0.70 |

The scatter plot shows how much one variable is affected by another and also the relationship between them. We have generated the graph as following.



We have displayed the relationship between the attributes using scatter plot. It helps us to find the relationship between x axis and y axis. In scatter plot each value in the data set is represented by a dotted x-axis represents prices, and the y-axis represents predicted prices. We can see that the dots are concentrated around the value 200000 on the x-axis, and 200000 on the y-axis. Using plt. show() function we have generated the graph. We can also see that the spread is wider in the lower middle part. Here using random forest we got the most accurate result-Squared is used to describe the relationship between the values of the x- and y-axis, and also to find the accuracy. The r-squared value ranges from 0 to 1, where 0 means no relationship, and 1 means 100% related.

**Code and output for the model in Random-forest regression:**



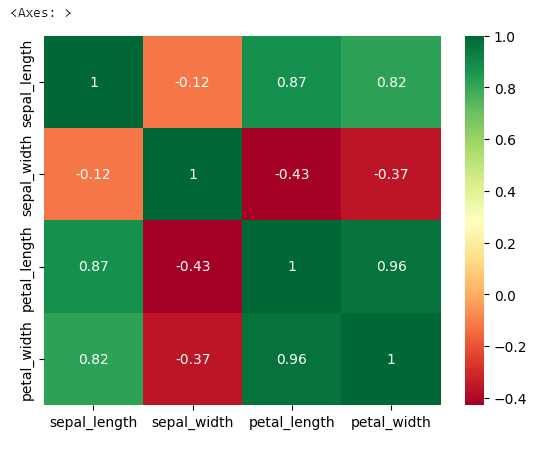
**Output prediction**:

R2 value-0.9547491075960973

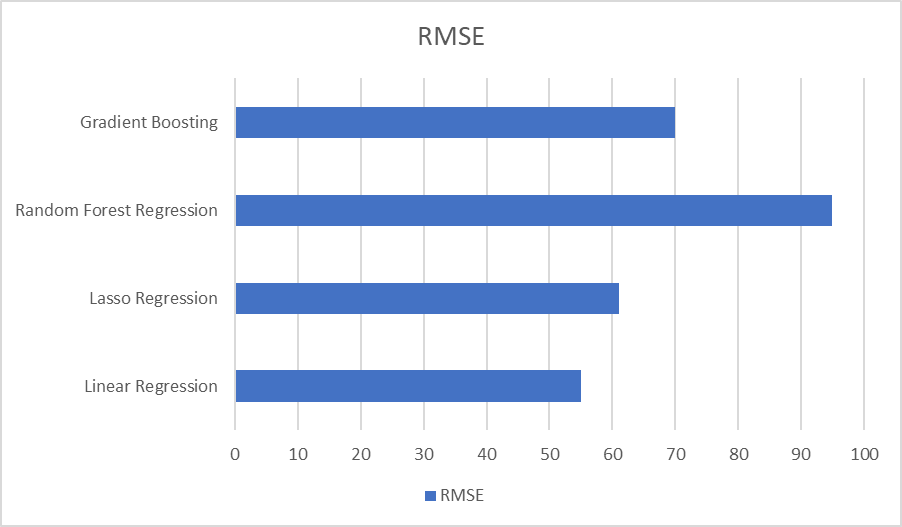
Adjusted R2- 0.9547431661568657

RMSE- 2100.81610381345

A heatmap is a visualization that shows a relationship between two dimensional matrix. In heatmap each value of a matrix is represented by different color patterns. We have also added the labels to represent each matrix. After predicting the code we have executed the heatmap we have got the following output.



Here is a bar graph to emphasis our result that we’ve found previously ->



**Chapter 6**

**6.1 Conclusion**

We’ve studied different research papers on the same. We add a few new attributes to predict better results than that paper. The city that we consider is one of the most important places in king county US or any nearest area. Because it is the best desirable place for living and here houses are 138% more expensive than the national average. Nearer to this place will be more expensive. After that, the preprocessed dataset was performed with different machine learning algorithms like the regression model. We got different results from different models. In the random forest regression algorithm, we got the most accurate result. 95% accuracy was found using a random forest regression algorithm.

**6.2 Future Work**

Though the model is capable enough to predict house prices, there is always a scope for improvement. In the future, this model will be enhanced to produce more accurate results. We will include the sentiment data and use it in the data pre-processing before testing and training experiments. Further, we can increase or update the data on a regular basis, to bring more accuracy to prediction. In our model, we can come up with an idea of using multiple functionalities along with the ML model. We can also use an Amazon EC2 connected with our model which will provide an easier user interface. Finally, our goal will be to develop a full-fledged web application that will predict house prices most accurately. We will also add a location feature in the application so the users can use it to predict the price of houses all over the world and at any location.